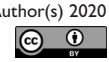
Special Issue: Against stored abstractions:
A radical exemplar model of language acquisition

FIRST LANGUAGE

First Language
2020, Vol. 40(5-6) 640–659



DOI: 10.1177/0142723720949723
journals.sagepub.com/home/fla

Abstractions made of exemplars or 'You're all right, and I've changed my mind': Response to commentators

Ben Ambridge
University of Liverpool, UK; ESRC International Centre for Language and Communicative Development (LuCiD), UK

Abstract
In this response to commentators, I agree with those who suggested that the distinction between exemplar- and abstraction-based accounts is something of a false dichotomy and therefore move to an abstractions-made-of-exemplars account under which (a) we store all the exemplars that we hear (subject to attention, decay, interference, etc.) but (b) in the service of language use, re-represent these exemplars at multiple levels of abstraction, as simulated by computational neural-network models such as BERT, ELMo and GPT-3. Whilst I maintain that traditional linguistic abstractions (e.g. a DETERMINER category; SUBJECT VERB OBJECT word order) are no more than human-readable approximations of the type of abstractions formed by both human and artificial multiple-layer networks, I express hope that the abstractions-made-of-exemplars position can point the way towards a truce in the language acquisition wars: We were all right all along, just focusing on different levels of abstraction.

Keywords
Abstractions, computational modelling, deep learning, exemplars, neural networks

I can't mean that, can I?

Pretty much! Let me explain. . . In the Before Times (I'm writing this under coronavirus lockdown) I wrote a target article (Ambridge, 2020) arguing that (a) we store every individual utterance that we hear, along with its understood meaning and contextual details, and (b) we do not store linguistic abstractions: apparent stored abstractions – such as those

Corresponding author:
Ben Ambridge, University of Liverpool, Eleanor Rathbone Building, Bedford St South, Liverpool L69 7ZA, UK.
Email: Ben.Ambridge@Liverpool.ac.uk

commonly posited to account for [SUBJECT] [VERB] [OBJECT] word order in English – are instead generated on the fly by analogizing across stored exemplars. Commentators wrote 18 replies – for which I am very grateful – many of them pointing out that the distinction between exemplar- and abstraction-based accounts is something of a false dichotomy (Demuth & Johnson, 2020; Finley, 2020; Lieven et al., 2020; McClelland, 2020; Mahowald et al., 2020; Rose, 2020; Schuler et al., 2020) And, do you know what? I was wrong and they were right. Of course, I don't agree with every detail. In particular, many of the phenomena raised by the commentators as evidence for stored abstractions strike me as equally compatible with *stored* abstractions and the *on-the-fly* abstractions that I argued for in the original target article. But let's not split hairs, since the whole point of the modified account that I sketch here is to collapse the exemplar–abstraction distinction.

The gist of this modified account is that, yes, we store all the exemplars that we hear (subject to attention, decay, interference, etc.), but that – in the service of language use – these exemplars are re-represented in such a way as to constitute abstractions (hence 'Abstractions made of exemplars'; see also Lieven et al.'s [2020] claim that 'children generalise at multiple levels of granularity'). As we will see in more detail shortly, a useful metaphor for this account is a multiple-level connectionist neural network that stores every exemplar, re-representing it in increasingly abstract ways as we move up the hidden layers (see also Blevins et al., 2018; and Li et al., 2007, as discussed by MacWhinney, 2020). Crucially, however, my claim is that this is not just a metaphor (Dennett, 2017; Hasson et al., 2020; Martin, 2020). The brain *really does* contain multiple layers of units (i.e. neurons), each of which aggregates input signals using a nonlinear function and outputs signals to other units. While any *particular* artificial neural network model of language is only the clumsiest metaphor, the claim that language is represented as patterns of activation across 'dumb' neurons, each of which 'knows' nothing about nouns, verbs and all the rest of it is literally true, and quite beyond dispute.[1]

Why the change of heart? Well, as several commentators (and journal reviewers) pointed out, my original target article – despite its title – provided much stronger evidence *for* stored exemplars than *against* stored abstractions. My argument against stored abstractions was essentially parsimony (if we have all the exemplars needed to generate abstractions on the fly, why do we need stored abstractions too?) plus the impossibility of positing abstractions that capture all of the data (e.g. both *John feared Bill* and *John frightened Bill* in the case of the English SVO transitive construction). But the point I was overlooking was this: if we store abstractions at multiple levels simultaneously, it doesn't matter if the highest-level abstractions don't explain every case; exemplars and lower-level abstractions are there to take up the slack.

A happy outcome of this abstractions-made-of-exemplars position is that it points the way towards a truce – or at least an armistice – in the language acquisition wars. We were *all* right all along, just focusing on different levels of abstraction. Chomskyans were impressed by speakers' abstractions at the highest levels – phrases, heads, complements and so on – dismissing low-level abstractions as 'just usage'. Advocates of chunk-based learning were impressed by speakers' abstractions at the lowest levels – individual *n*-grams like *What's+that*? and *cup+of+tea* – dismissing the highest-level abstractions as mere descriptive fictions with no psychological reality. Constructivists

were impressed by speakers' abstractions at the middle levels – by constructions like [AGENT] [VERB] [RECIPIENT] [THEME] – more abstract than mere lexical strings, but less abstract than phrases, heads and complements. We were all right. And yet, at the same time, all wrong: while the abstractions posited under the abstractions-made-of-exemplars account look *something* like phrases, lexical strings, argument-structure constructions and so on – and can be usefully summarized as such to a first approximation – they are in fact none of these things.

Can you tell me how to get, how to get to. . . abstract representations?

What *are* they? It's complicated, and I will try my best to explain below. But the whole point of this account is that the abstractions posited eschew verbal explanations. In fact, I would go so far as to say that demanding verbal explanations is where we have all been going wrong all of these years. For any system as complicated as language, it is naïve to expect an explanation couched in terms of 'human-readable' concepts like [NOUN] or [DATIVE CONSTRUCTION] to be anything more than a broad-brush sketch that should not be taken literally. To see why, let's use an analogy from a domain that is much closer than language to being 'solved': image classification (McClelland, 2020). Show a multi-level neural network model a picture, and it will tell you whether it's a cat, a dog or a house. How does it work? Well, if you insist on an explanation in terms of 'human-readable' concepts like 'nose', 'tail' and 'window' you can have one. But you know full well that this explanation is just a dumbed-down approximation generated to give humans some vague sense of how the system works. How does it actually work? The point is, nobody really knows; at least, not if you define 'knows' as 'able to give an explanation in terms of human-readable concepts'. The bottom-level, least-abstract layer represents the pixels of the image. As we move up through the layers, the representations become increasingly more abstract. If we plot the activation patterns of these more abstract layers and squint a bit, maybe we can just about make out something that looks sort of like a 'nose detector'. But we know full well that it isn't really one, and that any explanation couched in such simplistic terms is doomed to failure. Sorry, my fellow (psycho-/developmental-)linguists, but language is exactly the same.

We'll come back to language in a minute, I promise, but let's stick with image classification for just a moment longer because it nicely illustrates my central claim that multiple-layer neural networks – whether artificial or biological ones – are capable of storing both abstractions and a huge amount of exemplar-level information.[2] Until relatively recently, it was believed that image classification models succeeded only by forming abstractions that capture elements of many different exemplars (like our 'nose detector'). But Zhang et al. (2017) showed that an image-classification model can achieve perfect performance on the training set if the category labels are randomized, or even if each of the images is replaced by random pixels. Further evidence that my original dichotomy between exemplars and abstractions was misguided comes from Kelly et al. (2017), who showed that MINERVA 2 – a classic exemplar model – is mathematically equivalent to a particular type of distributed ('abstractions') model (a fourth order

tensor). As Demuth and Johnson (2020) rightly pointed out, 'feature-based approaches and exemplar-based approaches to learning are often formally equivalent'.

So now on to language, and with it the commentary by Mahowald et al. (2020). It should be clear by now that I wholeheartedly endorse their claim that 'there need not be a hard split between models that encode abstract structures and those that store a huge amount of information about the input and allow for fast analogical comparisons'. Indeed, this commentary – along with McClelland's (2020) and Schuler et al.'s (2020) – is what inspired my conversion. As these commentators point out, models like BERT (Devlin et al., 2018), ELMo (Peters et al., 2018) and GPT-2 (and, hot off the presses, GPT-3; Brown et al., 2020)[3] are capable of implicitly forming approximations to traditional 'syntactic categories' – and indeed of showing syntactic priming effects – while at the same time retaining the item-level information that governs the appropriate use of individual idiosyncratic members of these categories.

Of course, there is much to dislike about BERT and its ilk, primarily the fact that it lacks not only any kind of communicative goals, but any links to real-world meanings at all (Bender & Koller, 2020): Words are represented as vectors that capture their distributional similarity to other words (a kind of souped-up Latent Semantic Analysis), albeit in a context-dependent fashion (e.g. *table* would have different vectors in the input string *He sat at the table* and *See Table 1 for details*). And – let me be explicit – for this reason I am certainly not advocating BERT, or any other current model, as a feasible model of human language acquisition and representation. Until someone figures out how to implement communicative goals, discourse pragmatics, real-world meanings[4] and auditory rather than text-based representations – to name but a few – we're not even close. Furthermore, and perhaps most importantly, BERT's task – predicting the masked word in an input sentence – is nothing like the task facing human language learners; broadly speaking, to understand and to be understood. What I am advocating is the general architecture and approach; re-representing exemplars in a way that yields abstractions, while at the same time retaining a huge amount of exemplar-specific information. Just like human learners, BERT doesn't follow the traditional linguistic approach of positing only the most abstract representations possible; rather, it forms multiple abstractions at different grain sizes as a *by-product* of attempting to maximize its performance on some task.[5] (In this sense, it is the ultimate 'usage-based' model, although its usage goals are very different to those of humans.) As a multiple-layer network of units that re-represents linguistic input in increasingly abstract ways in the service of some goal, BERT allows us to think about language learning and representation in a way that goes far beyond paper-and-pencil linguistic theorizing and, at least in terms of its architecture, enjoys a degree of neurobiological plausibility.

While changing one's position is nothing to be ashamed of – we academics should try it more often – there is a sense in which it is rather unfair to commentators who were, of course, arguing with my old position, and not my new one. In the remainder of this response, as I revisit the domains covered in the original target article, I therefore consider the implications of each commentary not only for an *abstractions-made-of-exemplars* account, but for the pure *radical exemplar* account I originally advocated.[6]

But first, let me address the questions raised by Lieven et al. (2020) and Zettersten et al. (2020): What empirical findings would support and – most crucially – falsify my

current position? After all, if I'm now allowing for abstractions as well as exemplars, isn't any pattern of findings consistent with my position?

In fact, I don't think so. Although (contra Zettersten et al., 2020) I don't believe in *falsification* per se, even this new version of the account makes a testable prediction that could yield serious probabilistic Bayesian evidence against it. The prediction is this: If you probe speakers' linguistic knowledge in any domain, and if you use a sensitive enough test, you will *always* find the fingerprints of the original exemplars; any abstractions formed will never efface these exemplars completely. The most straightforward – but by no means only – 'fingerprints' are effects of surface frequency (see Ambridge et al., 2015, for other possibilities). For example, the account predicts that long after children become capable of applying English past tense *-ed* inflection in a wug test (Berko, 1958) – and, indeed, throughout their entire lifespan – they will produce quicker and more accurate responses for frequently witnessed forms, or for novel forms that are similar to them (e.g. Ambridge, 2010; Blything et al., 2017). For two more examples, see the experiments suggested by Hou and Morford (2020). Of course, *proving* a negative is impossible, which is why I don't like to talk about absolute falsification. But repeatedly finding evidence for a null effect (e.g. using Bayes Factors or frequentist equivalence testing), using a measure that is sensitive enough to detect observed effects in similar domains with the given sample size, would eventually constitute sufficient evidence to reject this proposal. This evidence would be particularly powerful if it were accompanied by effects of frequency (or whatever) at the level of the putative abstraction; a pattern that would constitute evidence that we store the abstraction, but not the exemplars.

Word meanings

I was surprised to see that not one of the commentators took issue with the central claim of this section of the original target article: that word meanings are structured as exemplars, rather than as prototype categories based around a central meaning. The reason I found this surprising is that the prototype view seems to me to be fairly well entrenched in the literature. This view is explicit in statements such as '*home* clearly has a prototypical meaning that can be extended to highlight a particular attribute (or attributes) of the prototypical meaning' (Goldberg, 2006, p. 169). It is implicit in a tremendous number of experimental studies of word-learning in which children must learn that a number of concrete objects with carefully controlled variations along – usually – a single feature are all instances of a *dax*. (I have yet to encounter a study in which children must also learn that entirely dissimilar concrete objects are also *daxes*, as are pictures of *daxes*, actions that are only abstractly related to the concrete objects, and so on; see my original discussion of the word *table*.)

So I don't think (contra Knabe & Vlach, 2020) that I was making a strawman argument. My perception of the word-learning literature (albeit as an outsider) is that the prototype view is the mainstream, textbook, 'common-sense' view, with the anti-representationalist positions adopted by researchers such as Linda Smith and Larissa Samuelson generally considered to be too radical. But I hope I'm wrong, as I'm basically in agreement with their position. Indeed, a number of studies conducted by these researchers (see Smith & Samuelson, 2006, for a commentary; see also Brooks &

Kempe, 2020) suggest that children's shape-bias effects need not be 'attributed to concepts as unitary representational entities' (p. 1342), calling into question the claim of Naigles (2020) that shape-bias effects can be used to draw conclusions regarding abstract categories.

Potentially problematic for my position are the findings discussed by Zettersten et al. (2020) showing that participants (a) rapidly generalize at superordinate levels (e.g. *vehicle*), (b) 'are more likely to ascribe a shared property (e.g., gills) to perceptually-dissimilar animals in the same biological category (e.g., sharks and tropical fish) than to perceptually-similar animals in different biological categories (e.g., sharks and dolphins) and (c) struggle with categorization rules that are complex or hard to verbalize. However, I am not entirely persuaded by these findings, given that they relate to categorization problems that seem to rely on fairly explicit verbalizable knowledge (e.g. that, despite their appearance, dolphins are mammals, not fish), rather than to naturalistic word-learning in young children (see also Brooks & Kempe, 2020 on the issue of explicitness).

None of this is to say that I reject the idea of word- and concept-level abstractions entirely, as I did in my original target article. The abstractions-made-of-exemplars position holds that individual word+meaning pairs are re-represented in increasingly abstract ways as we move up through the levels of the network. In some cases, at some levels, these representations may correspond *very approximately* to traditional prototype word meaning categories; but only in the sense that our image classification network can be said *very approximately* to have a 'nose detector'. Indeed, although they lack any *real* representation of word meanings, multi-layer neural network models such as BERT already work in something like this fashion. BERT's use of context-dependent vectors means that, initially, a word such as *table* will be represented by very different vectors in phrases such as *dining table, see Table 1 for details, league table* and *water table*. The model will form abstractions across distinct uses of *table* only when doing so aids it in its masked-word-prediction task.

Morphologically inflected words

With a couple of minor exceptions, nobody attempted to challenge my exemplar view of morphology. Of course, this may simply be because many of the commentators' primary interests lie elsewhere. I would like to think, however, that at least part of the reason for this omission is that the probabilistic frequency and phonological neighbourhood density effects that have been observed in this domain, and that are well simulated by exemplar models, are extremely difficult to explain under non-exemplar accounts. Hartshorne (2020) notes that symbolic 'models of morphology have pushed beyond 'single default rule' paradigms to employ myriad productive rules (O'Donnell, 2015)', but the model that he cites does not – as do the exemplar models I originally cited – make graded predictions regarding the acceptability of various past tense forms of novel verbs on the basis of their phonology; indeed, 'no phonological . . . structure is represented' at all (O'Donnell, 2015, p. 146).

MacWhinney (2020) is correct to point out that most of the exemplar models I cited in the target article lack a temporal dimension and so cannot simulate U-shaped learning in this domain (e.g. *went→goed→went*). However, this strikes me as a purely

implementational problem, since exemplar models are well placed to trade off effects of rote storage and analogy, both of which are their bread and butter (see for example the discriminative learning model of Ramscar et al. [2013], which in effect constitutes an exemplar model with a temporal element). Certainly, some production front-end would have to be added to simulate MacWhinney's horserace phenomenon (e.g. 'runned, uh ran'), but I see no reason why this would undermine the overall spirit of an exemplar+analogy model.

To my knowledge, nobody has yet attempted to extend exemplar models of the English past tense using a BERT-like architecture (the Kirov & Cotterell [2018] model cited in the original target article has some similarities, but has only two layers). Indeed, this relatively simple paradigm may not require such elaborate architecture. In principle, though, a BERT-type model is well placed to simulate acquisition of complex paradigms of inflectional morphology (such as the Polish verb system, modelled by Engelmann et al. [2019], using a connectionist network, though with only a hidden single layer). Because a BERT-type model posits commonalities only when doing so improves performance on its task (e.g. producing the requested person+number-marked present tense form given a verb stem), it retains the ability to maintain separate representations for infrequent or even entirely idiosyncratic mappings. At the same time, the abstract representations in the highest layers allow it to generalize to novel forms.

N-grams

Often, what commentators don't say is as informative as what they do. By my reckoning, just under half of the commentators seem to be of the view that our knowledge of language is best represented in terms of traditional high-level abstract categories with little verbatim storage of exemplars. Yet none of these commentators offered a suggestion of how such a model could explain our detailed knowledge of *n*-gram statistics. Why not? In my view, the answer is that such an explanation is close to impossible by definition. If we know (to borrow an example from Yang, 2013) that '*the bathroom* is more frequent than *a bathroom*' – and my target article cited considerable evidence that we do – we must, in some form or other, be storing these exemplars. If we throw them out entirely in favour of a DET+NOUN rule, there is simply no way to account for this detailed knowledge. Notably, the only commentary that addressed the issue of *n*-gram storage, Hou and Morford (2020), did so to extend the already-abundant evidence for chunking to signed languages, noting – as I did for spoken languages – the difficulty of accounting for such phenomena under accounts based solely on traditional high-level abstractions.

Sentence-level constructions (syntax)

Before considering some of the points raised by commentators in the domain of sentence-level syntax, I think it's valuable to once again point out what they did not say. In my target article, I spent six pages setting out why the notion of *as-abstract-as-possible* syntactic representations of (English) SUBJECT VERB OBJECT word order are both problematic in principle and difficult to reconcile with a great deal of data from experimental studies (mainly) with children. Yet none of the commentators who would

presumably advocate something like this notion provided any direct arguments against my claims, preferring – on the whole – to open up completely new lines of attack. Fair enough, maybe this isn't the time or place. But at some point, theorists of the view that speakers maintain only highly abstract representations of SVO word order (who can be found on both sides of the familiar generativist-constructivist divide) owe the field an answer to the question of how to reconcile their view with these findings. Notably, the only commentator who engaged directly with this issue, Chandler (2020), did so only to provide yet more evidence of exemplar storage at this level.

Adger (2020) asks how a radical exemplar model (REM) learner could learn that *Anson kept the picture in the shed* is ambiguous in a way that *Which shed did Anson keep the picture in?* is not. In one sense, Adger and I are starting from such different assumptions that it makes debate almost impossible. The key to his argument is that speakers learn to form questions by hearing declarative-question pairs such as *Jill liked the picture. ~ Which picture did Jill like?* and thus that when the declarative-question relationship in such pairs is analogized to the declarative *Anson kept the picture in the shed* to yield a question, it will bring along its baggage, in the form of its ambiguity. But his starting assumptions that speakers (a) learn questions from question+declarative pairs and thus (b) generate questions from corresponding declaratives are anathema not just to a radical exemplar model but to all usage-based accounts of question formation and acquisition (e.g. Ambridge & Rowland, 2009; Ambridge et al., 2006).

Yet, in another sense, Adger (2020) and I are almost in agreement. His broader point is that strings such as *the picture (that was) in the shed* and *my neighbour's cat's tail* can be used and understood correctly only if they are analysed as 'similar in abstract structure but different in surface form': i.e. as instances of something like a NOUN (or DETERMINER) PHRASE. I agree completely. As Demuth and Johnson (2020) put it, language-learning mechanisms 'will have to involve linguistic abstractions such as . . . syntactic phrases'. Of course. The only point of disagreement is where linguistic abstractions such as syntactic phrases come from. Is NOUN PHRASE a structure we're born with (which I take to be Adger's position), or one that emerges, in approximate form, either during on-the-fly analogy (my position in the original target article) or during increasingly abstract re-representation in a multiple-layer network pursuing some goal (my current position)? The latter options are on the table because nobody ever said that analogy operates solely on the basis of surface form. There is much else to go on. For example, *My neighbour, My neighbour's cat, My neighbour's cat's tail* and so on are similar in that they are potential (abstract) possessors,[7] and all are similar to *the picture that was in the shed* in that they are entities that can serve as topics, that can have properties attributed to them ('is huge'), and so on. Different in surface form, but similar in abstract (semantic, [discourse-]functional) structure.

For evidence that some approximation of a NOUN PHRASE can emerge during increasingly abstract re-representation in a multiple-layer network pursuing some goal, we need only try out a couple of sentences in BERT[8] (which, unlike human learners, *is* hobbled by a reliance on surface forms; albeit a lot of them).

(1) The picture in the shed [MASK] old

(2) The pictures in the shed [MASK] old

For (1), BERT predicts *is* with considerably greater probability than *are* (log probabilities −3.00 vs −11.48). For (2) BERT predicts *are* with considerably greater probability than *is* (−2.62 vs −9.29). The fact that BERT is not misled by the local bigram *shed+is* indicates that, while it does not have a NOUN PHRASE in so many words (remember our 'nose detector'), it has built some kind of abstract representation that effectively functions as one in this context (see Bernardy & Lappin, 2017; Linzen et al., 2016 for more systematic investigations).

If all three possibilities – innate NOUN PHRASE, on-the-fly approximation and BERT-style approximation – end up in more or less the same place, why should we discount the first option? Because if we retain only the abstraction, there is no way to account for the wealth of lower-level information that we in fact retain (e.g. that 'cat . . . can usually be preceded by an article' [Adger, 2020], but not by 'some', unless it is being served as a foodstuff; that *the+bathroom* is more frequent than *a+bathroom, cup+of+tea* than *cup+of+milk*, and so on).

I realize I've just spent a disproportionate chunk of my word allowance on a single commentary, but the investment is about to pay off, because the arguments I've just set out apply to all of the commentaries that apparently (to me at least) view syntax as dependent on stored, *as-abstract-as-possible* representations. For example, I certainly did not, as Hartshorne (2020) suggests, mean to imply, that speakers 'treat each utterance–meaning pair as its own exemplar, with no internal structure'; the meaning/function part of the pair *gives* the utterance internal structure (e.g. uniting *the picture that was in the shed* and *My neighbour's cat* as entities that can have properties attributed to them and so on). Similarly, phrases such as *pet vampire* are interpreted not by analogy with individual exemplars of pet and vampires, but with attributive noun phrases like *tomato soup*. Again, the meaning/function part of the pair – here something like [TYPE/PROPERTY][THING] – is what gives these utterances internal structure. Yet, again, we agree more than we disagree: yes, it is indeed 'difficult to imagine cognition without abstractions'; and, yes, they confer 'many blessings'. The point is that by having these abstractions emerge in the moment (as in the original target article) or as a by-product of re-representing exemplars in the service of some task (my current view) we can enjoy all the benefits of stored high-level abstractions (e.g. approximations of the SVO construction, NOUN PHRASE) with none of the costs that we incur if we posit *only* these abstractions (i.e. the failure to account for low-level lexical knowledge such as *n*-gram frequency).

I offer an almost identical rebuttal to Koring et al. (2020). It is absolutely not 'reasonable to expect a model of stored exemplars to assign the same meanings to the same strings of words' since – as I originally argued – 'learners store concrete exemplars, each including the surface form *along with its understood meaning and contextual details*'. A string of words such as 'Yeah, it was great!' can have understood meanings that are entirely opposite, given the contextual details (e.g. as responses to 'Did you enjoy your vacation?' [literal meaning] vs 'Did you enjoy your algebra exam?' [sarcasm]). Thus, the whole point of the exemplar approach is that each and every instance of a particular string (e.g. *the parents expected to like each other*) will be stored with slightly different understood meanings and contextual details. The same can be said for Koring et al.'s

(2020) other examples: the fact that *dis-* and *not* have similar interpretations in some contexts and dissimilar interpretations in others is a problem for a formalist model that attempts to shoehorn both into an abstract category of negation; not for an exemplar model that records every understood meaning and all contextual details for every *dis-* or *not* sentence it encounters. The fact that we say *did+not+agree+at+all* but *disagree+completely* is easily captured by even the simplest *n*-gram learner. Indeed, the native-speaker institution that *disagree+completely* is more common than *disagree+entirely* than *disagree+greatly* and so on is trivial for BERT,

(3) I have to say that I disagree [MASK].

completely (−1.77) > entirely (−3.25), greatly (−3.78), altogether (−3.93), absolutely (−4.38)

but a complete mystery for the 'formalist' models advocated by Koring et al. (2020).

Similarly, I agree with Messenger et al. (2020) that syntactic priming relies on some form of abstract syntactic structure (though I find it odd that they cite Bock [1989] and Bock & Loebell [1990] as evidence that 'priming occurs in the absence of lexical or thematic overlap' without challenging my claim in the target article that the findings of Ziegler & Snedeker [2018] and Ziegler et al. [2018], suggest otherwise; likewise, for counterevidence to Messenger et al., 2012; see Bidgood et al., in press). But, again, I see no advantage – and considerable disadvantage in terms of explaining low-level lexical effects – in assuming that this abstract syntactic structure is stored – and stored at only the most abstract level possible – as opposed to emergent (a) on the fly or (b) as a by-product of re-representing exemplars in the service of some task. Certainly, as Mahowald et al. (2020) point out, BERT-type models can simulate syntactic priming effects with only emergent knowledge of abstract syntactic structures. Crosslinguistic syntactic priming effects are, in effect, simulated by BERT's cousin Google Translate, which – when given an input sentence – produces an output sentence with corresponding syntax but different lexical items; the very definition of syntactic priming. This is easily demonstrated by, for example, entering English stimuli from Hartsuiker et al.'s (2004) seminal English–Spanish priming study, and inspecting the translations, which recapitulate the English syntax:

(4) The taxi chases the truck → El taxi persigue el camión

(5) The truck is chased by the taxi → El camión es perseguido por el taxi.

Nor am I persuaded by Messenger et al.'s (2020) argument that findings from amnesia patients 'are not well-explained by an exemplar model in which syntax generation is primarily influenced by the retrieval of exemplars, stored within declarative memory', given that I specifically advocated an exemplar model that 'blurs these [declarative/procedural, explicit/implicit] distinctions' (as, of course, do BERT-type models). I am, on the other hand, happy to concede that error-based (reverse-frequency) and timespan effects are not simulated by the kinds of exemplar models I originally discussed, which are mainly 'static' models without a time-course element. That said, I see nothing in principle that

prohibits exemplar models from adopting architectures that yield such effects; indeed, the BERT-type models that I advocate here use a form of error-based learning.

Naigles (2020) reports findings that children with ASD show correlations between their performance on early tasks measuring speed of processing SVO sentences and (1) later syntactic-bootstrapping tasks with very different SVO sentences and (2) later *wh*-question comprehension. Again, I would agree that the former finding requires *some* kind of appeal to abstract syntax, in that children's abilities with SVO at Time 1 are carried forward into Time 2. But, again, I see no reason to favour the version in which this abstract syntax takes the form of a stored SVO transitive construction at the most abstract level possible, as opposed to an approximation of an SVO transitive construction (and many less abstract sub-regularities) emergent (a) on the fly or (b) in BERT-style re-representation. Again, only the latter two options are compatible with the lexically restricted nature of young children's SVO transitive utterances as shown – for example – in the 19 production studies that I cited as evidence of an input-based advantage for pronoun-based over full-NP-based SVO transitives. The second finding, of correlations between performance on SVO transitives and *wh*-questions – and between each of these measures and vocabulary – is entirely expected under an account in which all rely on the same underlying ability of analogy (again, whether those analogies are conducted on the fly, or during BERT-style re-representation).

Phonetics and phonology

The arguments that I rehearsed in the previous section apply here too. Although many commentators pointed out the necessity of phonological abstractions, in general I struggled to understand the necessity of these abstractions being *stored*, rather than generated on the fly (of course, as Rose [2020] points out, '*blanket* rejection of abstract categories seems preposterous at best', which is why the original target article advocated only the rejection of *stored* abstractions). Fortunately, mainstream phoneticians and phonologists – on the whole, unlike syntacticians and morphologists – have preempted my belated conversion to a BERT-style model under which we both store individual exemplars and re-represent them into increasingly abstract representations that approximate – if only imperfectly – traditional categories. Indeed, although I would hardly describe myself as a fan of Optimality Theory in general, the 'multiple levels of abstraction' approach set out by Finley (2020) comes very close to the 'BERT-for-phonology' model that I would now advocate, as do the 'emergentist approaches to the development of increasingly abstract layers of categories based on the learner's past experiences' set out by Rose (2020).

The great strength of this approach is that, as I noted in the Introduction, it avoids positing a false dichotomy between 'feature-based approaches and exemplar-based approaches to learning [which] are often formally equivalent' (Demuth & Johnson, 2020; Kelly et al., 2017). In so doing, it retains all the advantages of exemplar approaches that I originally cited – retaining fine-grained information regarding speaker identity, sociolinguistic variation and so on – and acknowledges the fact that actual speech is far removed from the idealization of a sequence of phonemes drawn from a discrete inventory. It also retains all the advantages of abstraction-based approaches summarized by numerous commentators (see also McQueen et al., 2006; Turk & Shattuck-Hufnagel,

2020): we adapt quickly to unfamiliar accents (Hartshorne, 2020; MacWhinney, 2020; Schuler et al., 2020); infants in their first year generalize at the phonemic level, rapidly learning to ignore distinctions that are not relevant in the target language (Demuth & Johnson, 2020; Lieven et al., 2020; Zettersten et al., 2020); abstract phonotactic knowledge is necessary for word segmentation (Demuth & Johnson, 2020); although some fine-grained phonetic detail[9] of individual speakers is retained, much is lost (Schuler et al., 2020); orthographic representations are linked to idealized auditory representations of words rather than particular exemplars (Brooks & Kempe, 2020); speakers have intuitions about the relative acceptability of different consonant clusters, even if their language allows neither (Finley, 2020); learners' substitution patterns are not specific to individual word forms, but to segmental categories (Rose, 2020). In conclusion, in their widespread adoption of multiple-levels-of-abstraction accounts, developmental phonologists and phoneticians have shown us developmental syntacticians and morphologists a path that we would be wise to follow.

Bringing it all together

Before summing up, let me touch briefly – I promise – on a few important issues raised by the commentators that transcend the individual domains reviewed above.

Only a single commentary – Schuler et al. (2020) – directly addressed the neurological plausibility, or otherwise, of a radical exemplar model. I must admit that I lack the expertise needed to evaluate the findings raised by these commentators, but I am happy to take them on face value as evidence against my original radical exemplar model and in favour of a BERT-style model that represents both exemplars and abstractions at different neural levels.

Brooks and Kempe (2020) raise the issue of how explicit knowledge should be incorporated into an exemplar model. I'm sure explicit knowledge, particularly literacy, plays an important role, but I really have no idea how to account for it, other than to kick the can down the road: presumably, explicit knowledge, and certainly literacy, does not become an important factor until after the first few years of language acquisition, meaning those of us who are interested primarily in basic morphosyntax can largely afford not to worry about it. Later on, however, I'm sure Brooks and Kempe (2020) are entirely correct to point out that episodes of memory retrieval themselves create a memory trace. More generally, it is certainly the case that, as MacWhinney (2020) points out, the outputs of analogy need to be able to serve as the inputs to other analogies (including, for example, for possessive *'s* recursion; see note 7). Indeed, it would be unprincipled for an account that assumes that we store all the utterance–meaning pairs that we hear to make an exception for those that we produce ourselves (even if only internally).

Three commentators – Messenger et al. (2020), Knabe and Vlach (2020) and Brooks and Kempe (2020) – challenged my dismissal of forgetting-as-abstraction accounts. I agree that the issue is dealt with unsatisfactorily by classic exemplar accounts of the type I originally advocated, and – like with so much else in this response – see BERT-type models as the perfect way to bridge the gap. Such models store every exemplar (subject to attention, etc.) and do not 'forget' in the sense of expunging whole exemplars or their subparts. Yet, as Mahowald et al. (2020) point out, 'Due to various mechanisms

commonly used in neural networks such as regularization and incremental weight updating, typically not all training instances T are typically retrievable even if the number of parameters P > T.' In this respect, they simulate the all-too-human phenomenon of forgetting.[10] On the other hand, it is important to emphasize that, unlike the single-layer connectionist models I argued against in the original target article, such models nevertheless retain a huge amount of exemplar-level information (Zhang et al., 2017), allowing them – in principle – to simulate all of the exemplar-based phenomena I originally summarized.

The flipside of forgetting – sleep consolidation – was not explicitly raised by any of the commentators, but had been playing on my mind ever since the original target article. The phenomenon is that learning is most effective when it is distributed amongst different sessions with sleep in between (this includes children's learning of grammatical constructions, as per my very first published study; Ambridge et al., 2006); that is, memories formed during the day are somehow consolidated during sleep. This finding is difficult to square with a pure exemplar model, but is well simulated by BERT-style multiple-layer neural network models (e.g. Golden et al., 2020). Such models consolidate memories and avoid catastrophic forgetting (where new learning wipes out the old) using a 'sleep' period in which, just like in human sleep, patterns of neural firing from the pre-sleep period are replayed. It is precisely this period of offline reactivation that allows the model to form abstractions that – just like BERT's linguistic abstractions – apply across multiple exemplars.

Quite a few commentators – most explicitly Lieven et al. (2020), McClelland (2020), Demuth and Johnson (2020), Schuler et al. (2020) and Hou and Morford (2020) – raised the question of how to measure similarity (and whether this changes across time and across tasks); including the problems of how to segment the (linguistic and nonlinguistic) world into units over which similarity can be computed, and the extent to which different tokens of – at some level – 'the same' exemplar are really the same. All I can really say in response is that these questions apply equally to all accounts (a) on-the-fly analogy, (b) BERT-style abstractions made of exemplars, (c) high-level abstractions built from exemplars that are then discarded or (d) traditional linguistic categories with some innate basis. In all cases, learners need to know what dimensions are relevant, whether they are generalizing across exemplars to form abstract categories of some type, or simply assigning exemplars to them. Thus, the difficulty of specifying the generalization metric does not seem to me to favour one type of account over another.

Where do we go from here?

Pinker (1979, 1987) famously argued that sentences such as

John ate fish

John ate rabbits

John can fish

'would seduce a distributional analysis learner into combining heterogenous words . . . into a single class, leading to the production of' (for this example) **John can rabbits*, 'and other monstrosities' (Pinker, 1979, p. 240).[11] When I first started out in language acquisition research, this example was still being widely cited as evidence against the possibility of distributional analysis. These days, not so much. Why not? Well, in 1979, when distributional analysis was little more than a theoretical possibility, the argument was seductive. But as soon as researchers began to implement distributional analysis computationally, it turned out that it was not nearly so brittle (e.g. Mintz, 2003). Today, even the grammar-checker that is built into Microsoft Word – hardly the state of the art – baulks at **John can rabbits*.[12] Similarly, the question of how children learn to avoid errors such as **The clown laughed the man* (cf. *The clown made the man laugh*) while maintaining the ability to generalize unseen verbs into this construction has long been viewed as a 'learnability paradox' (Pinker, 1989, p. 415), and 'one of the most . . . difficult challenges for all students of language acquisition' (Bowerman, 1988, p. 73). But once we set aside the theoretical arguments and set out to build a computational model, it proved almost trivially simple (Ambridge et al., 2020). The English past tense debate – as discussed in the original target article – is a third example.

My point is that the argument 'It's impossible to learn X without innate knowledge/without abstract representations/using analogy' – or whatever – simply melts away once someone builds a computational model that does so. As Hartshorne (2020) puts it, 'We will not know what approach to modeling language acquisition will work until we have one that does' (see also Guest & Martin, 2020). Thus my first conclusion is that, if we are to move forward with this debate, we need much more computational modelling. In particular, we need *comparative* computational modelling: studies in which computational models instantiating the various theoretical approaches that have been advocated in this special issue and elsewhere are compared on their ability to closely simulate data from human children and adults, when supplied with comparable input.[13] After all, it's easy to find fault with any model, since a host of entirely unrealistic simplifying assumptions are necessary merely to implement the learning problem. This means it's not sufficient simply to find fault with Model X; you need to show that Model Y does a better job of simulating the human data. A difficulty here is that we have just about reached the point where building state-of-the-art computational models requires the resources of some of the world's biggest companies. If we mere academics are going to take on the might of Google, Uber and Elon Musk, we are going to have to pool our resources into a kind of CERN for language acquisition research[14] rather than – and I include myself entirely in this – tinkering around at the edges.

At the start of this commentary, I expressed the hope that an abstractions-made-of-exemplars position might point the way to a truce in the language acquisition wars. I would therefore like to end with my thoughts on what each side could learn from the other. But first, let me reiterate something else I said in the Introduction: concepts like *phoneme, word, determiner, phrase, construction* and so on are useful tools for thinking about language, but we should not kid ourselves that they are anything more than that. At the biological level, all of these things are no more than approximations of certain patterns of neural activity, and any explanation couched in terms of these human-readable concepts is nothing more than a broad-brush sketch.

Researchers on the usage-based-constructivist side should take more seriously the high-level, very abstract generalizations that have been uncovered by researchers on the generativist-nativist side. Under an abstractions-made-of-exemplars account, the fact that these generalizations often gloss over exceptions and facts of usage does not necessarily mean that they are wrong. Often, like the high-level abstractions of a BERT-type network, these high-level abstractions may merely need supplementing with much lower-level ones, sometimes all the way down to individual exemplar tokens. We may even be moving slowly towards a consensus on the vexed issue of innate knowledge. My impression – and I must admit, it is little more than that – is that generativist-nativist researchers are inching away from the position that categories and phrase structure themselves are innately given and towards the position that we are innately biased to form linguistic generalizations more readily on the basis of some properties than others (say, number than colour). Similarly, my impression is that many usage-based-constructivist researchers would have little difficulty in accepting these types of innate biases, particularly if they can be tied to some kind of universal hierarchy of communicative goals. For example, our model that solved Pinker's (1989, p. 415) 'learnability paradox' exemplified by **The clown laughed the man* (Ambridge et al., 2020) was set up in such a way as to form its generalizations on the basis of four measures of directness of causation. This was primarily done as an implementational convenience, but while we should be wary of evolutionary just-so stories, it would not seem particularly implausible for languages to evolve in such a way as to force their speakers to mark degrees of causality.

Conversely, researchers on the generativist-nativist side should take more seriously the low-level, very specific knowledge that has been uncovered by researchers on the usage-based-constructivist side. Under an abstractions-made-of-exemplars account, the fact that particular exemplars *could in principle* be united by higher-level abstractions does not necessarily mean that only the abstractions are represented, and that lower-level representations can be dismissed as learned exceptions or mere facts of usage. A satisfactory account of language acquisition, and indeed of adult representations, needs to account for our knowledge at all levels, from the most to the least abstract. Again, the solution may be simply to supplement the high-level abstractions with much lower-level ones, sometimes all the way down to individual exemplar tokens. And I see nothing in this that is incompatible in principle with the generativist-nativist position; or, at least, with the more moderate version that posits innate biases rather than innate categories.

In summary, we were all right and all wrong. At the descriptive level, language consists of both exemplars and abstractions at many levels of granularity, with different theoretical approaches preferring to zoom in on different levels. But at the implementational level, language consists of none of these things; just layers and layers of neurons.

Author contributions

Ben Ambridge: Conceptualization; Writing-original draft.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This project has received funding from the European Research

Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No. 681296: CLASS). Ben Ambridge is Professor in the International Centre for Language and Communicative Development (LuCiD) at the University of Liverpool. The support of the Economic and Social Research Council (ES/L008955/1) is gratefully acknowledged.

ORCID iD

Ben Ambridge https://orcid.org/0000-0003-2389-8477

Notes

1. Dupre and Yuste (2017) have already succeeded in decoding the behaviour of *Hydra Vulgaris* in terms of its neural circuits. It goes without saying that human language is considerably more complex than the behavioural repertoire of *Hydra Vulgaris* (essentially elongating and contracting), but unless one believes in 'wonder tissue' (Dennett, 2017) there is no reason to think that language, or any other human behaviour, cannot *in principle* be understood at this level.
2. As Mahowald et al. point out (see their note 2), 'typically not all training exemplars are typically retrieveable'; but, of course, the same is true for human language learners (see the section Bringing it all together).
3. For convenience, I will throughout this response talk about 'BERT-type models'. I don't intend this to mean just variations of BERT, or even just transformer models, but rather all models that use huge numbers of units and hidden layers to abstractly re-represent exemplars in the service of some task.
4. This seems to me to be the biggest stumbling block as it is unclear how it would ever be possible to add meanings at scale to a model. Even if we had the resources to hand-code every utterance, how could we ever agree on the semantic primitives to encode? (Of course, this is a problem for any computational model implementing any account of language, not just BERT.) For now, we will just have to live with context-dependent vectors as the least bad option. After all, while nobody will ever agree on the precise meaning of 'cat', we can all agree that it's closer in meaning to 'kitten' and 'dog' than to 'democracy' and 'whether'.
5. At least this is one rough-and-ready interpretation of what BERT-type models are doing. Exactly what they are doing is unclear (indeed, finding out is an active and fast-moving research area), but that's kind of the point: whatever they are doing is not readily understandable in terms of human-readable categories and concepts, and it is a mistake to try to reduce them to such.
6. Of course, I don't have the space to address each and every specific point made by every commentator – most made three or four separate points that could merit essay-length responses on their own – and for that I can only apologize.
7. For more detail see: www.lucid.ac.uk/news-events-blog/blogs/why-is-language-unlimited-david-adger-s-book-s-take-on-possessive-s/
8. I thank Kyle Mahowald for writing the code that allowed me to experiment with BERT's masked sentence prediction task.
9. As I neglected to make clear in the original target article, my use of the term 'phonetic detail' (and indeed of 'phonetics' and 'phonology') is intended to include signed languages.
10. Interestingly, Khandelwal et al. (2020) and Kassner and Schütze (2020) show that augmenting a BERT-type model with exemplars (specifically a *k*-nearest neighbours model) boosts performance, precisely because it allows for storage of infrequent patterns that are otherwise 'forgotten'; i.e. erased by 'mechanisms . . . such as regularization and incremental weight updating' (Mahowald et al., 2020). Since human learners do seem to forget infrequent

patterns, the goals of maximizing model performance and simulating human behaviour may not be aligned here.

11. '. . . The words are coming out all weird.'
12. Word's autocorrect suggests 'rabbit' meaning something like 'talk incessantly, despite the fact that your listener is clearly bored'; a meaning that I suspect is specific to British English (apparently from Cockney rhyming slang, 'rabbit and pork', 'talk'; e.g. https://en.wikipedia.org/wiki/Rabbit_(song)).
13. As Guest and Martin (2020, p. 3) put it, 'All models are wrong but some are more wrong than others (pastiche based on: Box, 1976; Orwell, 1945).'
14. In fact, there already exist several large-scale international collaborations with ambitious goals to map and simulate the human brain, such as the United States BRAIN Initiative and the European Union Human Brain Project (see Yuste & Bargmann, 2017, for a summary). But as far as I could make out, language is barely touched upon in either.

References

Adger, D. (2020). Syntax and the failure of analogical generalisation: A commentary on Ambridge (2020). *First Language 40*(5-6): 560–563.

Ambridge, B. (2010). Children's judgments of regular and irregular novel past tense forms: New data on the dual- versus single-route debate. *Developmental Psychology*, *46*(6), 1497–1504.

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language 40*(5-6): 509–559.

Ambridge, B., Maitreyee, R., Tatsumi, T., Doherty, L., Zicherman, S., Mateo-Pedro, P., Bannard, C., Samanta, S., McCauley, S., Arnon, I., Bekman, D., Efrati, A., Berman, R., Narasimhan, B., Sharma, D. M., Nair, R. B., Fukumura, K., Campbell, S., Pye, C., . . . Mendoza, M. J. (2020). The crosslinguistic acquisition of causative sentence structure: Computational modeling and grammaticality judgments from adult and child speakers of English, Japanese, Hindi, Hebrew and Kiche. *Cognition*, *202*, Article 104310. https://doi.org/10.1016/j.cognition.2020

Ambridge, B., & Rowland, C. F. (2009). Predicting children's errors with negative questions: Testing a schema-combination account. *Cognitive Linguistics*, *20*(2), 225–266.

Ambridge, B., Rowland, C. F., Theakston, A., & Tomasello, M. (2006). Comparing different accounts of non-inversion errors in children's non-subject Wh-questions: 'What experimental data can tell us?' *Journal of Child Language*, *30*(3), 519–557.

Ambridge, B., Rowland, C. F., Theakston, A. L., & Kidd, E. J. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*(2), 239–273.

Ambridge, B., Theakston, A., Lieven, E. V. M., & Tomasello, M. (2006). The distributed learning effect for children's acquisition of an abstract grammatical construction. *Cognitive Development*, *21*, 174–193.

Bender, E. M., & Koller, A. (2020, July). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics.

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*(2–3), 150–177.

Bernardy, J. P., & Lappin, S. (2017). Using deep neural networks to learn syntactic agreement. *Linguistic Issues in Language Technology*, *15*, 1–15.

Bidgood, A., Rowland, C. F., Pine, J. M., & Ambridge, B. (in press). Syntactic representations are both abstract and semantically constrained: Evidence from children's and adults' comprehension and production/priming of the English passive. *Cognitive Science*.

Blevins, T., Levy, O., & Zettlemoyer, L. (2018, July). Deep RNNs encode soft hierarchical syntax. In I. Gurevych, & Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the*

Association for Computational Linguistics (Vol. 2, Short papers) (pp. 14–19). Association for Computational Linguistics.

Blything, R. P., Ambridge, B., & Lieven, E. V. M. (2017). Children's acquisition of the English past-Tense: Evidence for a single-route account from novel verb production data. *Cognitive Science*, *42*(S2), 621–639.

Bock, K. (1989). Closed-class immanence in sentence production. *Cognition*, *31*(2), 163–186.

Bock, K., & Loebell, H. (1990). Framing sentences. *Cognition*, *35*, 1–39.

Bowerman, M. (1988). The 'no negative evidence' problem: How do children avoid constructing an overly general grammar? In J. A. Hawkins (Ed.), *Explaining language universals* (pp. 73–101). Blackwell.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799.

Brooks, P. J., & Kempe, V. (2020). How are exemplar representations transformed by encoding, retrieval, and explicit knowledge? A commentary on Ambridge (2020). *First Language 40*(5-6): 564–568.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language learners are few-shot learners. arXiv preprint. https://arxiv.org/abs/2005.14165.

Chandler, S. (2020). Sentence-level constructions: A demonstration in support of Ambridge (2020). *First Language 40*(5-6): 569–572.

Demuth, K., & Johnson, M. (2020). Exemplar-based learning probably requires learning abstractions: A commentary on Ambridge (2020). *First Language 40*(5-6): 573–575.

Dennett, D. C. (2017). *From bacteria to Bach and back: The evolution of minds*. Norton.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dupre, C., & Yuste, R. (2017). Non-overlapping neural networks in Hydra vulgaris. *Current Biology*, *27*(8), 1085–1097.

Engelmann, F., Granlund, S., Kolak, J., Zreder, M., Ambridge, B., Pine, J. M., Theakston, A. L., & Lieven, E. V. M. (2019). How the input shapes the acquisition of verb morphology: Elicited production and computational modelling in two highly inflected languages. *Cognitive Psychology*, *110*, 30–69.

Finley, S. (2020). The need for abstraction in phonology: A commentary on Ambridge (2020). *First Language 40*(5-6): 576–580.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Golden, R., Delanois, J. E., Sanda, P., & Bazhenov, M. (2020). Sleep prevents catastrophic forgetting in spiking neural networks by forming joint synaptic weight representations. *bioRxiv* 688622.

Guest, O., & Martin, A. E. (2020). *How computational modeling can force theory building in psychological science*. https://psyarxiv.com/rybh9/download?format=pdf

Hartshorne, J. (2020). The many blessings of abstraction: A commentary on Ambridge (2020). *First Language 40*(5-6): 581–584.

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? *Psychological Science*, *15*(6), 409–414.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, *105*(3), 416–434.

Hou, L., & Morford, J. (2020). Using signed language collocations to investigate acquisition. A commentary on Ambridge (2020). *First Language 40*(5-6): 585–591.

Kassner, N., & Schütze, H. (2020). BERT-kNN: Adding a kNN search component to pretrained language models for better QA. arXiv preprint arXiv:2005.00766.

Kelly, M. A., Mewhort, D. J., & West, R. L. (2017). The memory tesseract: Mathematical equivalence between composite and separate storage memory models. *Journal of Mathematical Psychology*, *77*, 142–155.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2020, April 28–May 1). *Generalization through memorization: Nearest neighbor language models* [Conference session]. International Conference on Learning Representations (ICLR).

Kirov, C., & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, *6*, 651–665.

Knabe, M. L., & Vlach, H. A. (2020). Anti-representationalism in language development research: A commentary on Ambridge (2020). *First Language 40*(5-6): 592–595.

Koring, L., Giblin, I., Thornton, R., & Crain, S. (2020). Like dishwashing detergents, all analogies are not the same: A commentary on Ambridge (2020). *First Language 40*(5-6): 596–599.

Li, P., Zhao, X., & MacWhinney, B. (2007). Self-organizing processes in early lexical learning. *Cognitive Science*, *31*, 581–612.

Lieven, E., Ferry, A., Theakston, A., & Twomey, K. E. (2020). Similarity, analogy and development in radical exemplar theory: A commentary on Ambridge (2020). *First Language 40*(5-6): 600–603.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

MacWhinney, B. (2020). The role of competition and timeframes: A commentary on Ambridge (2020). *First Language 40*(5-6): 604–607.

Mahowald, K., Kachergis, G., & Frank, M. C. (2020). What counts as an exemplar model, anyway? A commentary on Ambridge (2020). *First Language 40*(5-6): 608–611.

Martin, A. E. (2020). A compositional neural architecture for language. *Journal of Cognitive Neuroscience*, *32*, 1407–1427.

McClelland, J. (2020). Exemplar models are useful and deep neural networks overcome their limitations: A commentary on Ambridge (2020). *First Language 40*(5-6): 612–615.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.

Messenger, K., Branigan, H. P., McLean, J. F., & Sorace, A. (2012). Is young children's passive syntax semantically constrained? Evidence from syntactic priming. *Journal of Memory and Language*, *66*(4), 568–587.

Messenger, K., Hardy, S. M., & Coumel, M. (2020). An exemplar model should be able to explain all syntactic priming phenomena: A commentary on Ambridge (2020). *First Language 40*(5-6): 616–620.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91–117.

Naigles, L. R. (2020). Atypical language development matters: A commentary on Ambridge (2020). *First Language 40*(5-6): 621–625.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Orwell, G. (1945). *Animal farm: A fairy story*. Secker and Warburg.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human language technologies (Volume 1, Long Papers) (pp. 2227–2237). Association for Computational Linguistics.

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*(3), 217–283.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 399–441). Routledge.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT Press.

Ramscar, M., Dye, M., & McCauley, S. M. (2013). Error and expectation in language learning: The curious absence of 'mouses' in adult speech. *Language*, *89*(4), 760–793.

Rose, Y. (2020). There is no phonology without abstract categories: A commentary on Ambridge (2020). *First Language 40*(5-6): 626–630.

Schuler, K. D., Kodner, J., & Caplan, S. (2020). Abstractions are good for brains and machines: A commentary on Ambridge (2020). *First Language 40*(5-6): 631–635.

Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology*, *42*(6), 1339–1343.

Turk, A., & Shattuck-Hufnagel, S. (2020). Timing evidence for symbolic phonological representations and phonology-extrinsic timing in speech production. *Frontiers in Psychology*, *10*, Article 2952.

Yang, C. (2013, December). Who's afraid of George Kingsley Zipf? *Significance*, pp. 29–34.

Yuste, R., & Bargmann, C. (2017). Toward a global BRAIN initiative. *Cell*, *168*(6), 956–959.

Zettersten, M., Schonberg, C., & Lupyan, G. (2020). What does a radical exemplar view not predict? A commentary on Ambridge (2020). *First Language 40*(5-6): 636–639.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017, April 24–26). *Understanding deep learning requires rethinking generalization* [Conference session]. 5th International Conference on Learning Representations (ICLR 2017), Toulon, France.

Ziegler, J., Goldberg, A., & Snedeker, J. (2018, March). *Passive priming requires function word overlap* [Poster presentation]. 31st Annual Meeting of the CUNY Conference on Human Sentence Processing, Davis, CA.

Ziegler, J., & Snedeker, J. (2018). How broad are thematic roles? Evidence from structural priming. *Cognition*, *179*, 221–240.